



**Using ChemAxon Technology for
Computer Curation of patents & the scientific literature**



Patents contain molecular data in multiple forms :



Text – Image – manually created chemicals

As bitmap images

Pictures of chemicals found in the document Images

US 7,504,509 B2

113
diluted with ethyl acetate (100 ml). The layers were separated and the aqueous layer was extracted with ethyl acetate (4x, 75 ml). The organic layers were combined and washed once with brine (100 ml) before drying over sodium sulfate. The organic layer was evaporated under reduced pressure to yield the titled product as an off-white solid, (52 mg, M=1, 161.2)

Example 2
Synthesis of 3-(3-Methoxy-benzyl)-5-thiophen-3-yl-pyrrolo[2,3-b]pyridine 14 and 3-(5-Thiophen-3-yl-1H-pyrrolo[2,3-b]pyridine-3-ylmethyl)-phenol 15

Scheme 45

114
-continued

Step-1 Synthesis of 5-Bromo-1H-pyrrolo[2,3-b]pyridine-3-ylmethyl-dimethyl-amine 13.
Into a Round bottom flask was added 5-bromo-7-azaindole (54.0 mg, 0.002741 mol) and Dimethylamine hydrochloride (0.44 g, 0.0030 mol) and Paraformaldehyde (0.090 g, 0.0030 mol) and Isopropyl alcohol (40.0 mL, 0.522 mol). The reaction mixture was heated reflux for 17 hours. The reaction mixture was poured into water, followed by adding K_2CO_3 till pH 9. Then the aqueous layer was extracted with EtOAc. The organic layer was washed with brine, dried over sodium sulfate, concentrated and purified with biotage to give product 13 381.0 mg, together with 180.0 mg starting material recovered.

Step-2 Synthesis of 5-Bromo-3-dimethylaminomethyl-pyrrolo[2,3-b]pyridine-1-carboxylic acid tert-butyl ester 16.
Into a Round bottom flask was added compound 13 (380.0 mg, 0.001495 mol) and N,N-Dimethylformamide (10.0 mL, 0.119 mol) and sodium hydride (66 mg, 0.0016 mol). 10 minutes later, was added t-Bu-tert-butylcarbamate (650 mg, 0.0030 mol). The reaction mixture was stirred at room temperature for another 2 hours. TLC indicated no starting material. The reaction mixture was poured into water, extracted with EtOAc. The organic layer was washed with brine, dried over sodium sulfate, concentrated and dried with oil pump over weekend to give 540 mg product 16.

Step-3 Synthesis of 3-Dimethylaminomethyl-5-thiophen-3-yl-pyrrolo[2,3-b]pyridine-1-carboxylic acid tert-butyl ester 17.
Into a Round bottom flask compound 16 (628.0 mg, 0.001773 mol) and 3-thiophene benzoic acid (390.0 mg, 0.00348 mol) and Potassium carbonate (300.0 mg, 0.00738 mol) and Tetrakis(triphenylphosphine)palladium(0) (40.0 mg, 0.000346 mol) and Tetrahydrofuran (16.0 mL, 0.197 mol) and Water (4.0 mL, 0.22 mol) under an atmosphere of Nitrogen. The reaction was heated to reflux overnight. The reaction mixture was poured into water, extracted with EtOAc. The organic layer was washed with brine, dried over sodium sulfate, concentrated and purified with biotage to give product 17 (600.0 mg).

Step-4 3-Chloromethyl-5-thiophen-3-yl-pyrrolo[2,3-b]pyridine-1-carboxylic acid tert-butyl ester 18.
Into a Round bottom flask was added compound 17 (120.0 mg, 0.00034 mol) and Toluene (4.0 mL, 0.035 mol) under an atmosphere of Nitrogen. To the reaction mixture was added Ethyl chloroacetate (40.0 mg, 0.00034 mol). The reaction

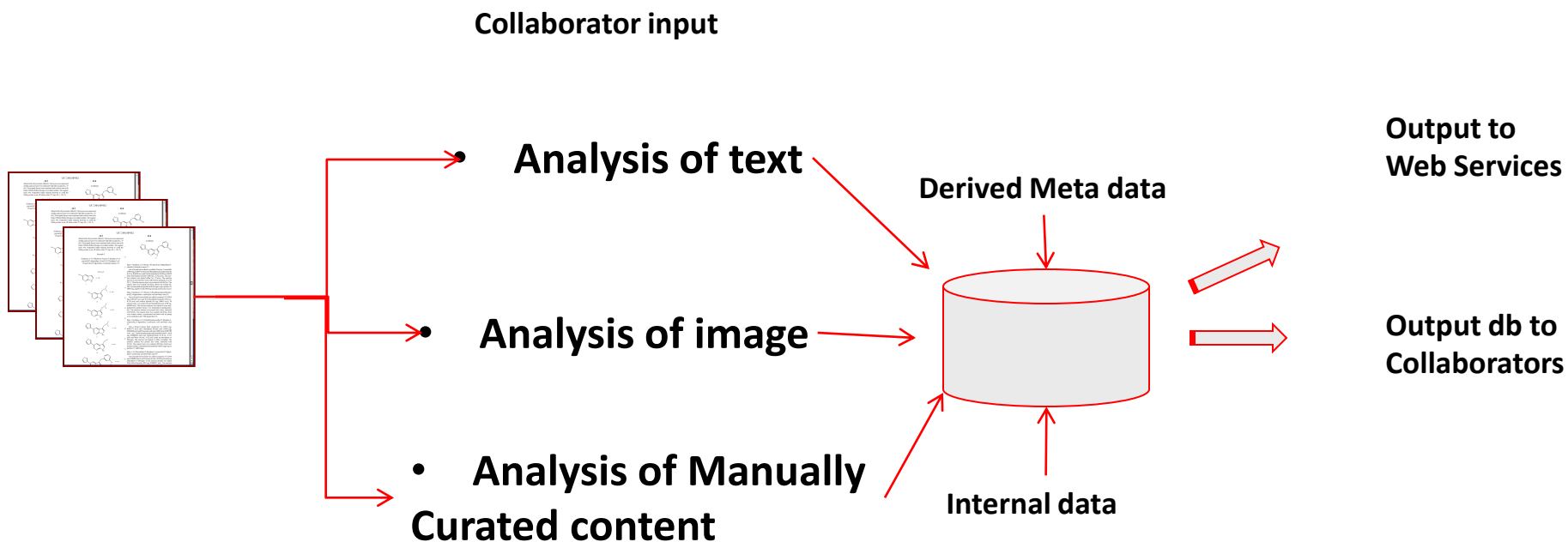
As text

Chemical names found in the text of documents

Computer curation now involves multiple types of analysis



combining technologies into workflow protocols



Can you find the key molecule's in an unstructured text , for example a scientific journal or patent?



Chemical nomenclature can be daunting

a) (2P/4S)-4-[4-Amino-5-(4-benzyloxy-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-2-hydroxymethyl-pyrrolidine-1-carboxylic acid tert-butyl ester prepared analogously to Example 18 starting from (2R/4S)-4-[4-amino-5-(4-benzyloxy-phenyl)-pyrrolo[2,3-d]pyrimidin-7-yl]-pyrrolidine-1,2-dicarboxylic acid 1-tert-butyl ester 2-ethyl ester (Example 20a). ¹H-NMR (CDCl₃, ppm): 8.52 (s, 1H), 7.52-7.32 (m, 7H), 7.1 (d, 2H), 6.95 (d, 1H), 5.50 (m, 1H), 5.13 (s, 2H), 4.62-4.42 (m, 2H), 4.28 (m, 2H), 4.10 (m, 1H), 3.95-3.70 (m, 1H), 2.75 (m, 1H), 2.50 (m, 1H), 1.49 (s, 9H).

b) (2R/4S)-{4-[4-Amino-5-(4-benzyloxy-phenyl)-pyrrolo[2,3-d]pyrimidin-7-yl]-pyrrolidin-2-yl}-methanol: 0.100 g of (2R/4S)-4-[4-amino-5-(4-benzyloxy-phenyl)-pyrrolo[2,3-d]pyrimidin-7-yl]-pyrrolidine-1,2-dicarboxylic acid 1-tert-butyl ester is dissolved in 4 ml of tetrahydrofuran; 10 ml of 4M hydrogen chloride in diethyl ether are added, and stirring is carried out for 1 hour at room temperature. The product is filtered off and dried under a high vacuum. The dihydrochloride of the title compound is obtained. ¹H-NMR (CD₃OD, ppm): 8.4 (s, 1H); 7.60 (s, 1H), 7.5-7.10 (m, 9H), 5.65 (m, 1H), 5.18 (s, 2H), 4.32 (m, 1H), 4.00-3.65 (m, 4H), 2.60 (m, 2H).

EXAMPLE 24

(2R/4S)-4-(4-Amino-5-phenyl-pyrrolo[2,3-d]pyrimidin-7-yl)-1-(2,2-dimethyl-propionyl)-pyrrolidine-2-carboxylic acid ethyl ester 0.130 g of (2R/4S)-4-(4-benzyloxycarbonylamino-5-phenyl-pyrrolo[2,3-d]pyrimidin-7-yl)-1-(2,2-dimethyl-propionyl)-pyrrolidine-2-carboxylic acid ethyl ester is dissolved in 8 ml of methanol, and the solution is hydrogenated over 0.030 g of palladium-on-carbon (10%) for 1 hour at normal pressure. The catalyst is removed by filtration, the filtrate is concentrated by

identify the chemical names – then convert them to structures
[chemical names -> structures] !



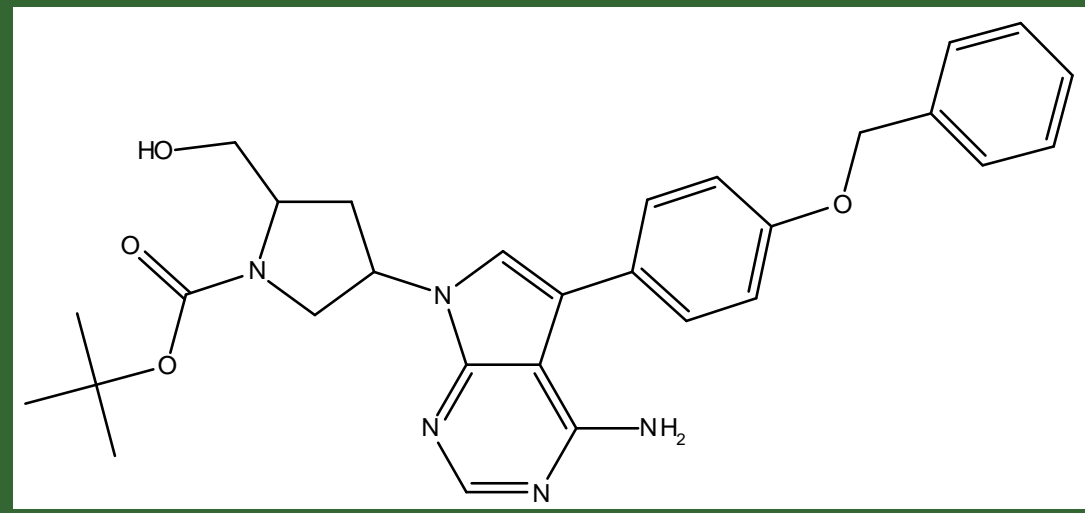
entity identification

a) (2R/4S)-4-[4-Amino-5-(4-benzyloxy-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-2-hydroxymethyl-pyrrolidine-1-carboxylic acid tert-butyl ester prepared analogously to Example 18 starting from (2R/4S)-4-[4-amino-5-(4-benzyloxy-phenyl)-pyrrolo[2,3-d]pyrimidin-7-yl]-pyrrolidine-1,2-dicarboxylic acid 1-tert-butyl ester 2-ethyl ester (Example 20a). $^1\text{H NMR}$ (CDCl_3 , ppm): 8.52 (s, 1H), 7.52-7.32 (m, 7H), 7.1 (d, 2H), 6.95 (d, 1H), 5.50 (m, 1H), 5.13 (s, 2H), 4.10 (m, 2H), 3.95-3.70 (m, 1H), 2.75 (m, 1H),

b) (2R/4S)-{4-[4-amino-5-(4-benzyloxy-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-2-hydroxymethyl-pyrrolidine-1-carboxylic acid tert-butyl ester} prepared analogously to Example 18 starting from (2R/4S)-4-[4-amino-5-(4-benzyloxy-phenyl)-pyrrolo[2,3-d]pyrimidin-7-yl]-pyrrolidine-1,2-dicarboxylic acid 1-tert-butyl ester 2-ethyl ester (Example 20a). 0.100 g of (2R/4S)-4-[4-amino-5-(4-benzyloxy-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-pyrrolidine-1,2-dicarboxylic acid 1-tert-butyl ester 2-ethyl ester in diethyl ether was filtered off and dried. $^1\text{H NMR}$ (CD_3OD , ppm): 8.52 (s, 1H), 7.52-7.32 (m, 7H), 7.1 (d, 2H), 6.95 (d, 1H), 5.50 (m, 1H), 5.13 (s, 2H), 4.10 (m, 2H), 3.95-3.70 (m, 1H), 2.75 (m, 1H),

(2R/4S)-4-[4-amino-5-(4-benzyloxy-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-2-hydroxymethyl-pyrrolidine-1-carboxylic acid tert-butyl ester prepared analogously to Example 18 starting from (2R/4S)-4-[4-amino-5-(4-benzyloxy-phenyl)-pyrrolo[2,3-d]pyrimidin-7-yl]-pyrrolidine-1,2-dicarboxylic acid 1-tert-butyl ester 2-ethyl ester in diethyl ether, filtered off and dried. $^1\text{H NMR}$ (CD_3OD , ppm): 8.52 (s, 1H), 7.52-7.32 (m, 7H), 7.1 (d, 2H), 6.95 (d, 1H), 5.50 (m, 1H), 5.13 (s, 2H), 4.10 (m, 2H), 3.95-3.70 (m, 1H), 2.75 (m, 1H),

What is this compound ??



anol:
2-
bromide
is
1 H-
H),
2-
ml of
at

identify the chemical names – then convert them to structures
[chemical names -> structures] !



entity identification

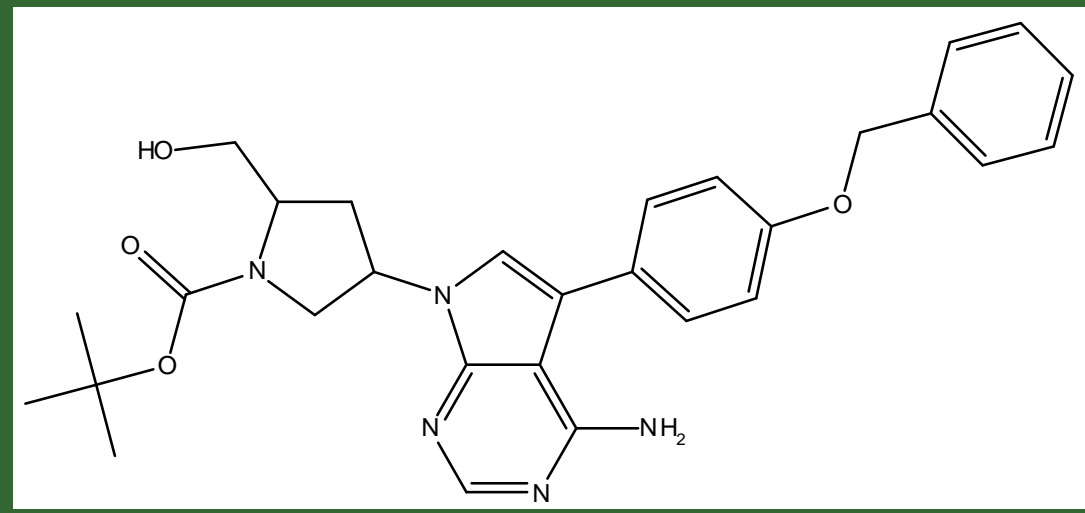
a) (2R/4S)-4-[4-Amino-5-(4-benzyloxy-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-2-hydroxymethyl-pyrrolidine-1-carboxylic acid tert-butyl ester prepared analogously to Example 18 starting from (2R/4S)-4-[4-amino-5-(4-benzyloxy-phenyl)-pyrrolo[2,3-d]pyrimidin-7-yl]-pyrrolidine-1,2-dicarboxylic acid 1-tert-butyl ester 2-ethyl ester (Example 20a). $^1\text{H NMR}$ (CDCl_3 , ppm): 8.52 (s, 1H), 7.52-7.32 (m, 7H), 7.1 (d, 2H), 6.95 (d, 1H), 5.50 (m, 1H), 5.13 (s, 2H), 4.10 (m, 2H), 3.95-3.70 (m, 1H), 2.75 (m, 1H),

b) (2R/4S)-{4-[4-amino-5-(4-benzyloxy-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-2-hydroxymethyl-pyrrolidine-1-carboxylic acid tert-butyl ester} in diethyl ether, filtered off and dried. $^1\text{H NMR}$ (CD_3OD) (ppm): 8.52 (s, 1H), 7.52-7.32 (m, 7H), 7.1 (d, 2H), 6.95 (d, 1H), 5.50 (m, 1H), 5.13 (s, 2H), 4.10 (m, 2H), 3.95-3.70 (m, 1H), 2.75 (m, 1H).

(2R/4S)-4-[4-amino-5-(4-benzyloxy-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-2-hydroxymethyl-pyrrolidine-1-carboxylic acid tert-butyl ester

methanol, and dried. Normalized

What is this compound ??



ChemAxon [Name = Structure]

Leading Causes of Annotator Problems *



Typical problems encountered when dealing with OCR text

Improper spacing within the chemical name:

2- (Bicyclo [2.2. 1] hept-5-en-2-ylamino) -5- [2- (4-chloro-3-methylphenoxy) ethyl]-l, 3- thiazol-4 (5H)-one

Run on lists:



indane, 1,2, 3,4- tetrahydroquinoline, 3, 4-dihydro-2H-1, 4-benzoxazine, 1,5-naphthyridine, 1, 8- naphthyridine

Numbering of compounds:



Comparative Example 3, 2-bromo-4- (1, 3-dioxo-1, 3-dihydro-2H-isoindol-2-yl) butanoic acid 4-(1,3-dioxo-1,3-dihydro-2H-isoindol-2-yl) butanoic acid

Formatting issues:

2-[2-(bicyclo [2.2. 1] hept-5-en-2-ylamino) -4-oxo-4, 5-dihydro-1, 3-thiazol-5-yl] -N-

 (4-metlioxyphenyl)-N-methylacetamide

Missing or Incorrect Parenthesis:



5-(2-anilinoethyl)-2-[(2-cyclohex-1-en-1-ylethyl)amino]-1,3-thiazol-4(5H)-one



* using WO/2005/075471 as an example

ChemAxon [Structure Checker]

What about processing image data ??



Image entity recognition

IBM pioneered a process for converting images of chemical structures – into Mol files (machine readable representations of chemical structures...)

We can also analyze the image content of patents & journals

Seminal paper on converting chemical images into MOL files

Optical Recognition of Chemical Structures (OROCS)

Richard Casey, Stephen Boyer, Paul Healey, Alex Miller, Bernadette Oudot, and Karl Zilles

IBM Almaden Research Center, 650 Harry Road, San Jose, California 95120 USA

Abstract

A prototype system for encoding chemical structure diagrams from scanned printed documents is described. The system distinguishes a structure diagram from other printed material on a page image using size and spacing characteristics. It distinguishes line graphics from symbols in an intermediate vectorization stage. Line information is mapped into a connection diagram that represents atomic bonds. Atomic symbols are identified by means of chemical drawing conventions and optical character recognition. The final coded output interfaces to conventional chemistry software for database storage and retrieval, publishing, and modeling.

Keywords: document analysis, graphics recognition, line drawing analysis, pattern recognition, chemical graphics

Introduction

Today, there are vast databases of chemical and biological information, all dependent on graphical representations of molecules as the critical feature allowing data to be accessed via substructure searching techniques. Once a database is created, it serves as the central facility for a wealth of other applications, such as information retrieval, publishing, scientific analysis, etc (see Figure 1). Facilities for entering graphical data are less advanced than those for manipulating it. For many years this problem has impeded the transfer to computers of paper systems such as utility maps, engineering diagrams, graphical chemical data etc.. To create a graphical object in digital format, an engineering diagram for example, requires appreciable time on the part of a trained operator. Frequently, it requires a duplication of effort in the sense that the operator works from an existing drawing or hand sketch. Chemical structures that are candidates for addition to databases, for example, are often already printed in journals and catalogs.

In this paper we describe an investigation into the creation of a coded representation of a chemical structure starting from an optically scanned diagram on a printed page. This is a pattern recognition process that calls for a preliminary analysis of the input document in order to discriminate a chemical structure diagram from other printed information on the page. Then follows a sequence of

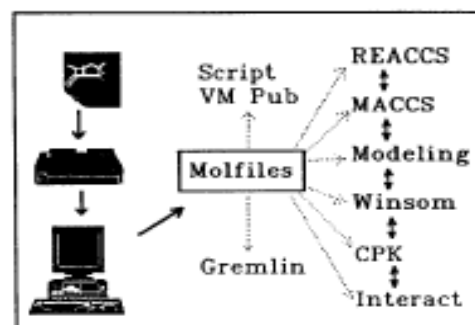


Figure 1. MOL-file applications.

steps that detects lines and determines their interrelations, recognizes geometric shapes, distinguishes printed characters and encodes them by means of OCR, and creates a connection table expressing the results. This must be done while accommodating different drawing conventions, and, importantly, applying the rules of chemistry to resolve ambiguities and to validate results. It is the problem of encoding engineering drawings, but on a smaller scale, and with a wealth of contextual information available to be applied in the process.

Since this project was begun in 1988, other researchers have reported on the recognition of chemical structures using different techniques [1-4]. The methodology reported here was granted a United States patent in 1992 [5].

Representation of Chemical Structures

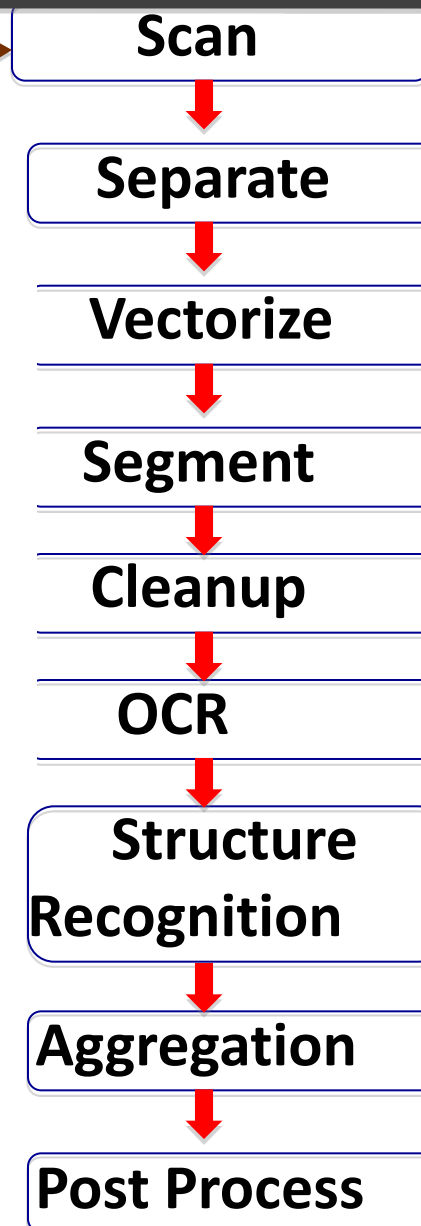
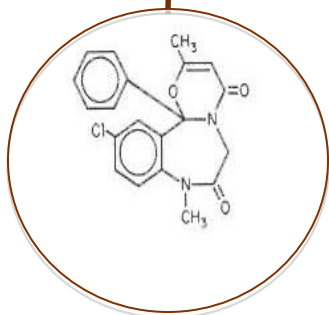
A number of equivalent methods are available for representing the chemical structure of a molecule, e.g., adjacency matrices, connection tables, and linked lists. Recent efforts to develop standards for molecular connection tables range from the Brookhaven Protein Data Bank format to the Molecular Design Limited (MDL) MOLFILE format, the Standard Molecular Data (SMD) format [6] and others [7]. For our purposes, we have elected to use the MDL MOLFILE [8] as the format for the graphical representation of the chemical structures for our optical recogni-

Optical recognition of chemical structures (OROCs)

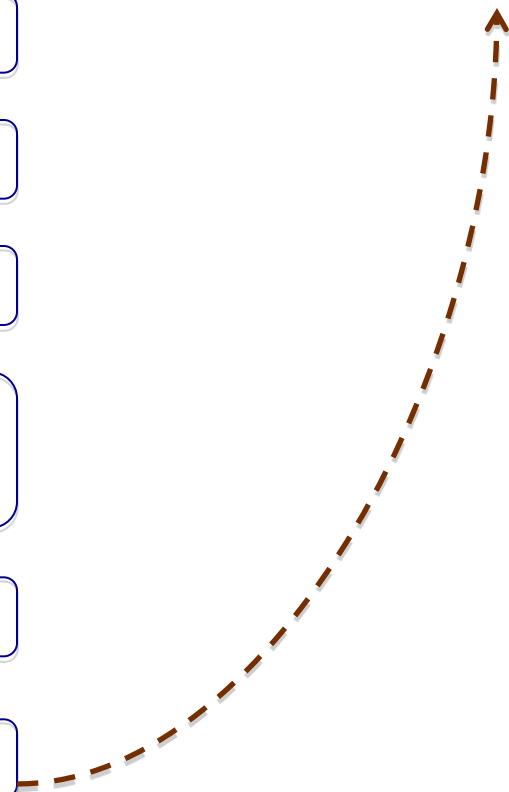
– How it works



Ketazolam [1971] (kee tay' zoe lam). C₂₀H₁₇ClN₂O₃. 368.81. (1) 4*H*-[1,3]-Oxazino[3,2-*d*][1,4]benzodiazepine-4,7(6*H*)-dione, 11-chloro-8,12*b*-dihydro-2,8-dimethyl-; (2) 11-Chloro-8,12*b*-dihydro-2,8-dimethyl-12*b*-phenyl-4*H*-[1,3]-oxazino[3,2-*d*][1,4]benzodiazepine-4,7(6*H*)dione. CAS-27223-35-4. INN; BAN. *Tranquilizer (minor)*. ◇U-28,774



O=C(CN1C2(C3=CC=CC=C3)OC(C)=CC1=O)N(C)C4=C2C=C(Cl)C=C4



Today there are numerous programs to do this :



Chemical image recognition programs

Programs for image recognition of chemical structures

- OSRA - (NIH)
- Clide Pro (Keymodule ltd - Peter Johnson)
- Kukleae
- Fraunhofer chemoCR (Fraunhofer Inst)
- ChemReader - (Univ of Michigan)

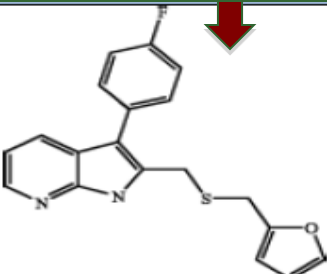
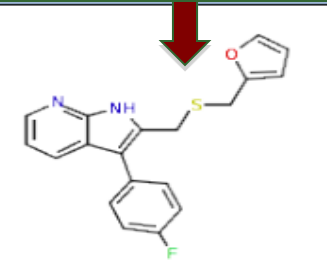
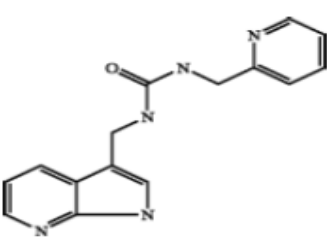
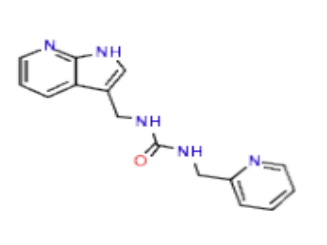
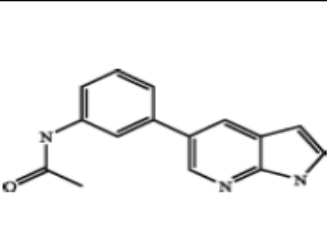
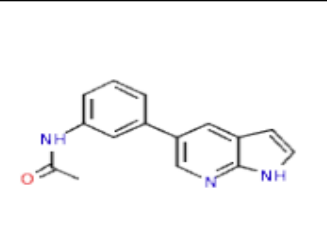
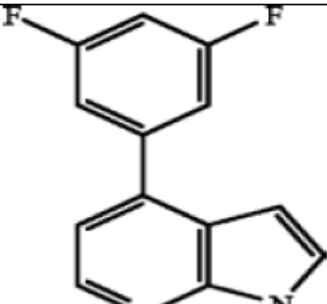
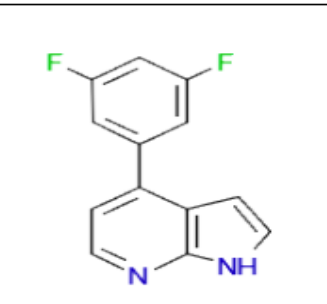
Image
Extracted from
the page

Structure
Generated from
the image

SMILE String
Generated from
the image



Chemical derived from OCR of image data

| | | |
|--|---|---|
|  |  | <chem>Fc1ccc(cc1)c1c(CSCc2ccoco2)[nH]c2ncccc12</chem> |
|  |  | <chem>O=C(NCc1ccccc1)NCc1c[nH]c2ncccc12</chem> |
|  |  | <chem>CC(=O)Nc1ccccc1c1cnc2[nH]ccc2c1</chem> |
|  |  | <chem>Fc1cc(F)cc(c1)c1ccnc2[nH]ccc12</chem> |

Examples :
Results from
OCR
of chemical
images

Image
Extracted from
the page

Structure
Generated from
the image

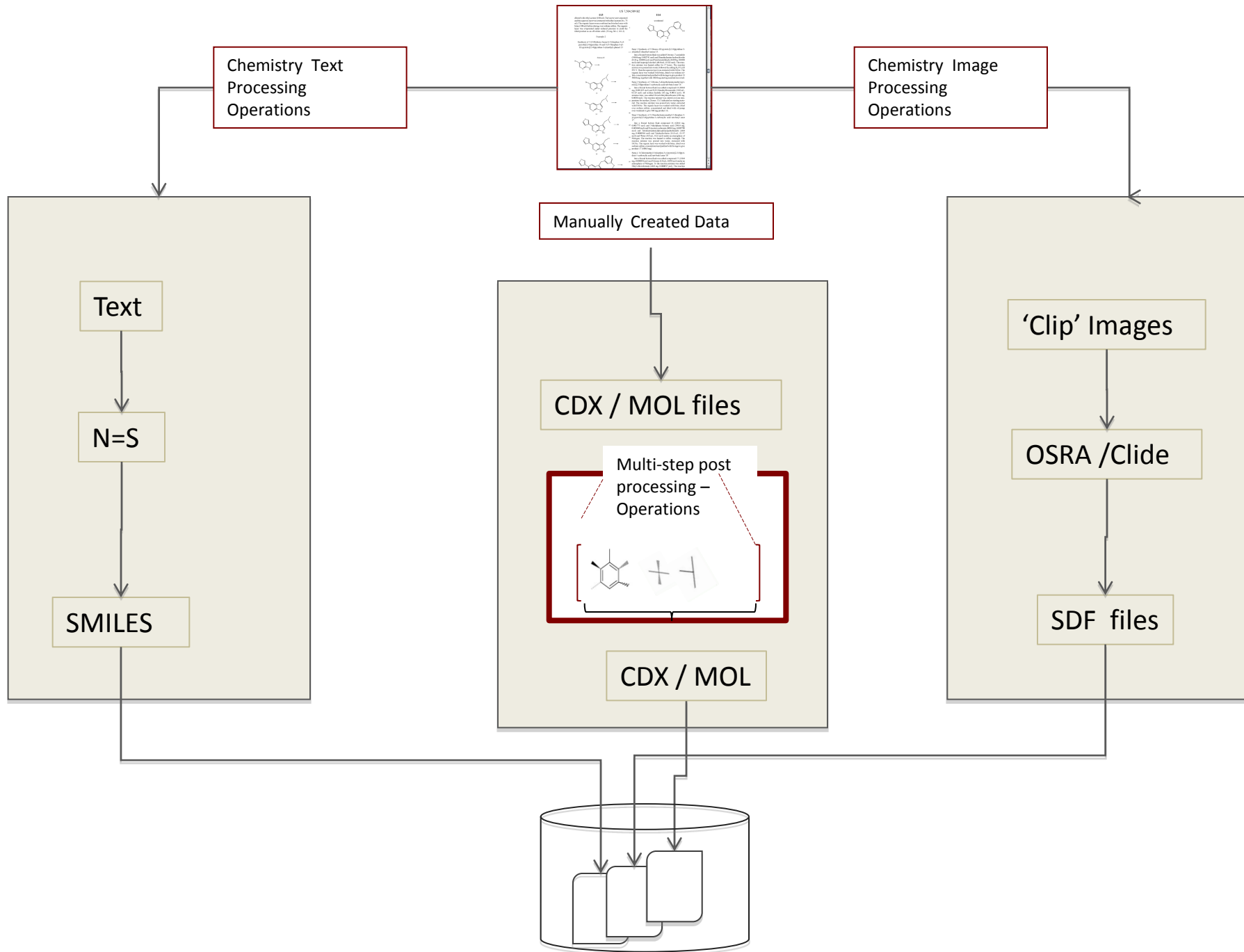
SMILE String
Generated from
the image

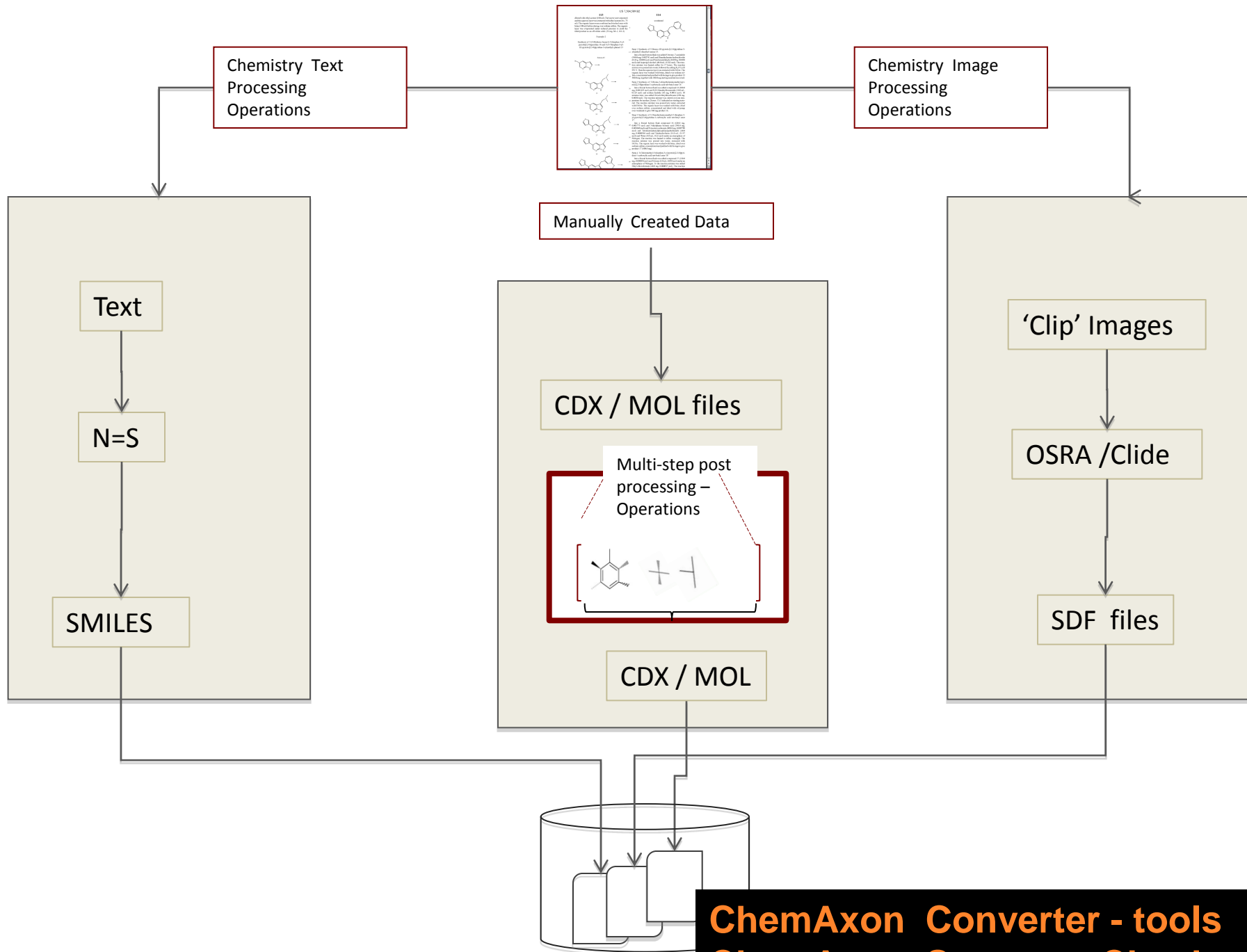


Chemical derived from OCR of image data

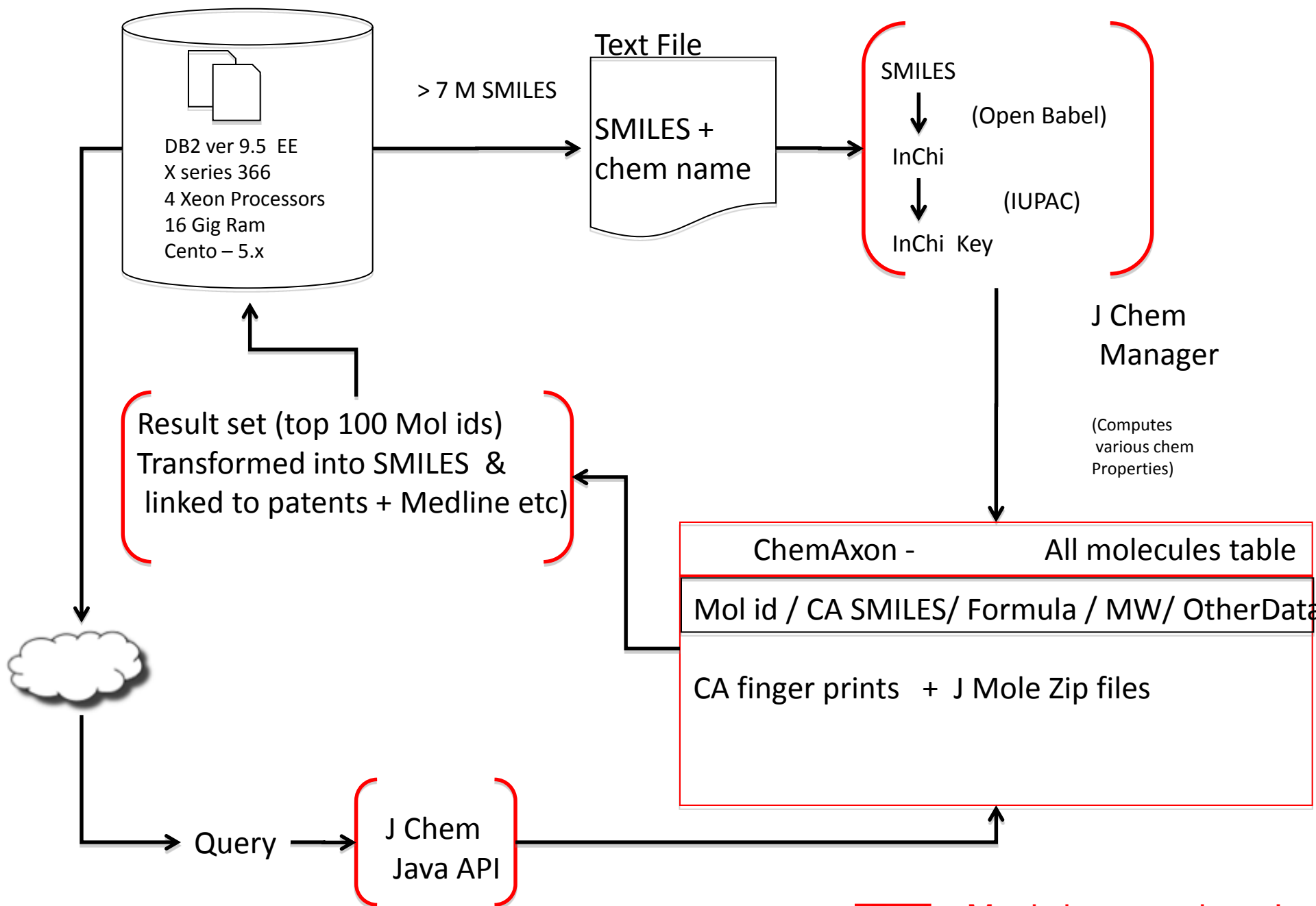
| | | |
|--|--|---|
| | | <chem>Oc1cccc(c1)C(=O)c1c[nH]c2ncc(cc12)c1cccs1</chem> |
| | | <chem>Oc1cccc(c1)C(=O)c1c[nH]c2nccc(c3cc(F)cc(F)c3)c12</chem> |
| | | <chem>Cc1onc(c1)C(=O)c1c[nH]c2ncc(cc12)c1cccnc1</chem> |
| | | <chem>Cc1onc(c1)C(=O)c1c[nH]c2ncc(cc12)c1cscs1</chem> |

Examples :
Results from
OCR
of chemical
images



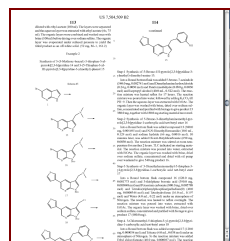


ChemAxon Converter - tools
ChemAxon Structure Checker



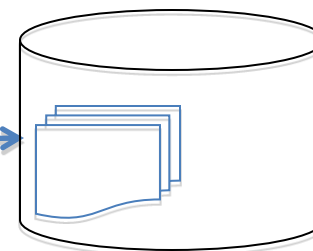
Red = Magic happens here !

ChemAxon Technologies that enable computer curation



- **Name = Structure**
- **Structure = Name**
- **Structure Checker**
- **Marvin View**
- **J Chem Manager**
- **J Chem**
- **Instant J Chem**

Magic happens here



SIMPLE